DOCUMENT RESUME

ED 215 564 FL 012 837

AUTHOR Sankoff, David; Poplack, Shana
TITLE A Formal Grammar for Code-Switching. CENTRO Working Papers 8.
INSTITUTION City Univ. of New York, N.Y. Centro de Estudios Puertorriguenos.
PUB DATE 80
NOTE 58p.; Prepared by Language Policy Task Force. For related document see ED 214 394.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Bilingualism; *Code Switching (Language); English; Language Research; Language Usage; Puerto Ricans; Spanish; Surface Structure; *Syntax; Vocabulary

ABSTRACT

This study, part of an on-going investigation, analyzes the syntactic aspects of code-switching. A series of empirical studies has confirmed that there are only two general linguistic constraints where code-switching may occur, the free morpheme constraint and the equivalence constraint. This study describes in formal terms how the two constraints determine the way two monolingual grammars may be combined in generating discourse containing code-switches. The data were selected from recordings of 20 individuals including both balanced bilinguals and speakers who are fluent in Spanish but not in English. The data, already analyzed in a previous study (Poplack, 1979), were reanalyzed within a formal grammatical framework based on the direct generation of surface phrase structures by a context-free grammar. The analysis sketches surface grammars for the Spanish and English spoken in the corpus, as well as for the code-switching mode. The analysis also studies the syntactic aspects of code-switching heard in the speech of Puerto Rican bilinguals. One point that emerged is that code-switching constraints are surface phenomena and that phrase structure grammars for each language can be combined to form a code-switching grammar which generates grammatical monolingual and code-switching sentences. (AMH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED215564

CENTRO

WORKING PAPERS

FL012837

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Center for Puerto Rican Studies

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Centro de Estudios Puertorriqueños

A FORMAL GRAMMAR FOR CODE-SWITCHING

8

Language Policy Task Force

Copyright ©, RESEARCH FOUNDATION OF
THE CITY UNIVERSITY OF NEW YORK, 1980

APRENDER A LUCHAR

LUCHAR ES APRENDER

CENTRO DE ESTUDIOS PUERTORRIQUENOS 445 W. 59th St.; New York, N.Y. 10019 Tel.: (212) 489-5260

A FORMAL GRAMMAR FOR CODE-SWITCHING[1]

David Sankoff
Centre de recherches mathématiques
Université de Montréal
and
Shana Poplack
Center for Puerto Rican Studies, CUNY

1. Introduction

Among the diverse configurations of linguistic performance in communities where two or more languages are in contact, the alternating use of different languages within a given situation, or code-switching, is a well-documented pattern. Much progress has been made in situating code-switching within a micro-sociological framework or that of the ethnography of speaking, consistent with the goals of understanding the interactive purpose, communicative function and social implications of this behavior (e.g. Gumperz 1964, 1971, 1976; G. Sankoff 1968, 1972; Denison 1972; Gumperz and Hernández-Chavez 1970; McClure and Wentz 1975; McClure 1977; di Sciullo et al. 1976; Valdés-Fallis 1976, 1978). A relatively small number of studies have focused directly on the grammatical aspects of code-switching (e.g. Hasselmo 1972, 1979; Gingràs 1974; Lance 1975; Timm 1975, 1978; Pfaff 1975, 1976, 1979; Wentz 1977; Lipski 1978).

Complete understanding of code-switching could only be achieved through combined ethnographic, attitudinal and grammatical study, i.e. an integrated analysis not only of when people code-switch, but

how, where and why. The present paper is but part of such an ongoing investigation; though here we concentrate on the purely syntactic aspects of code-switching, we in no way minimize the social determinants and implications of this behavior, which previous reports have explored in conjunction with the linguistic aspects (Poplack 1978, 1979a, 1979b).

We distinguish code-switching from other possible outcomes of language contact situations such as interference, pidginization, borrowing, calquing, language death, relexification, learned use of foreign words, cross-language punning and other word-play, by at least two criteria. One is that whereas many of the above involve deformation or replacement of parts of the grammar or lexicon of the language(s) involved, code-switching does not. This is one of the basic postulates of this paper. Second, unlike other of the above-mentioned phenomena, which refer to specialized situations or language functions, what we understand by 'code-switching' here is a widely operative norm of communication in certain types of multilingual communities (see also G. Sankoff 1972; Pedraza ms.). These characteristics of code-switching--the structural integrity of the component languages, and its prevalence in a broad range of communicative situations--have deep implications for grammatical theory. Insofar as discourse is generally thought of as being generated through the coherent pragmatic, semantic and syntactic mechanisms of a language shared by members of a community, how can two distinct languages

reconcile their differences in such a way as to result in discourse involving language switches not only between utterances, but also within a single sentence? More specifically, how can we construct a formal account of the grammatical mechanism which underlies discourse containing code-switching?

Note that there is no syntactic difficulty involved in alternating whole sentences, or larger segments, of different languages as in (1); this practice is common among bilinguals responding to a change in interlocutor, topic or setting (e.g. Weinreich 1953; Gal 1978).

(1) ¿Tu eres ateo? ¿Tu eres ateo? [You're an atheist?]
No he's not. He believes in something. (C.A./44)

The real problem involves the maintenance of syntactic integrity of a single sentence containing elements of two or more languages, as in (2):

(2) So you *todavía* haven't decided *lo que vas a hacer* next week. [So you still haven't decided what you're going to do next week.] (P.A./135)

A series of empirical studies of verbal interaction in one of the oldest Puerto Rican communities in the United States (Poplack 1978, 1979a, 1979b) has confirmed that there are only two general linguistic constraints on where switching may occur:

a) The free morpheme constraint: a switch may not occur between a bound morpheme and a lexical form unless the latter has been phonologically integrated into the language of the bound morpheme.

This excludes switches like (3), in which the phonology of *run* is unambiguously English, while that of *eando* is unambiguously Spanish (and which in fact do not occur), but not forms like (4). Indeed, we consider here phonologically, morphologically and syntactically integrated items like the latter to be Spanish forms, and not instances of code-switching.

(3) *run - *eando* [ɹʌn-e'ando] 'running'

(4) *flipeando* [flipe'ando] 'flipping'

b) The equivalence constraint: the order of sentence constituents immediately adjacent to and on both sides of the switch point must be grammatical with respect to both languages involved simultaneously. This requires some specification: the local co-grammaticality or equivalence of the two languages in the vicinity of the switch holds as long as the order of any two sentence elements, one before and one after the switch point, is not excluded in either language.

The equivalence constraint is illustrated in Figure 1, where the dotted lines indicate permissible switch points, and the arrows indicate the surface relationship of the two languages. Switches may occur at, but not between, the dotted lines.

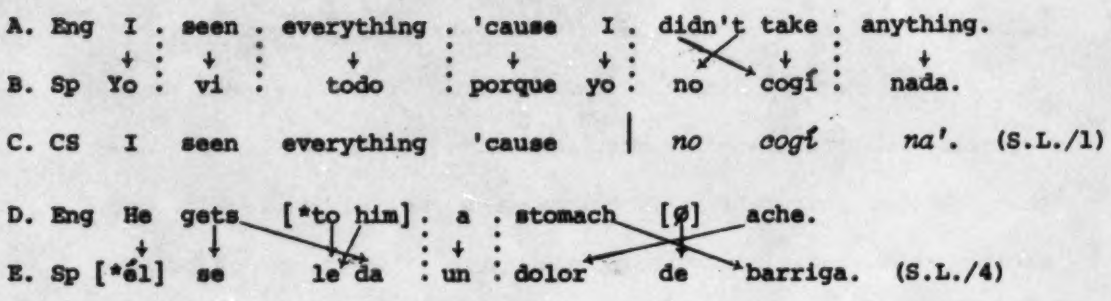

A. Eng I . seen . everything . 'cause I . didn't take . anything.
B. Sp Yo : vi : todo : porque yo : no cogí : nada.
C. CS I seen everything 'cause | no cogí na'. (S.L./1)
D. Eng He gets [*to him] . a . stomach [ø] ache.
E. Sp [*él] se le da : un : dolor de barriga. (S.L./4)

Figure 1. Permissible code-switching points. The speaker's actual performance is represented in (C), containing one switch, and (E), containing no switch.

Linguistic performance constrained in this way must be based on simultaneous access to the grammatical rules of both languages. This raises the question of the existence and nature of a code-switching grammar. In this paper we describe in formal terms how the code-switching constraints determine the way the two monolingual grammars may be combined in generating discourse containing code-switches.

Aside from its purely formal interest, this analysis will illustrate how code-switching, especially intra-sentential code-switching, rather than representing a debasement of linguistic skills, as certain prescriptivists claim (e.g. de Granda 1968; Varo 1971; LaFontaine 1975), is a development requiring competence in the two component codes, as well as the additional skill to manipulate them concurrently.

2. Sociolinguistic context and the universality of constraints

Concurrent with the enunciation of the two general code-switching constraints--free morpheme and equivalence--it was shown that the more particularistic constraints posited previously, for example that single determiners or subject pronouns cannot be switched (Timm 1975; Gumperz 1976; Wentz 1977), or any other such restrictions, are not borne out empirically, except where they are consequences of the two general constraints.

However, establishing the status of the free morpheme and equivalence constraints as universal or near-universal conditions on switching would require much comparative empirical work. Aside from the Puerto Rican data, they have been verified for Chicano materials published by Valdés-Fallis (1976) and Pfaff (1975, 1976), Swedish-English (Hasselmo, personal communication), and Italian-French-English (di Sciullo et al. 1976) code-switching, and in a preliminary though quantitative way on Greek-English, French-English, Italian-English and Yiddish-Spanish-Hebrew data (D. Tong, and S. Papadopoulos, D. Sheeh, F. Marchese and D. Litvak, New York University class papers).

However, it is not clear how the free morpheme constraint might operate in a situation involving English and some highly inflected or agglutinative language, nor what might be the scope of the equivalence constraint for languages with highly different word orders. To be pertinent, evidence in such cases would depend on

establishment of rigorous criteria for (a) distinguishing switches from borrowing, calquing or relexification patterns which may have become part of the monolingual norm, (b) identifying possible equivalence constraint violations against a background of information on monolingual word order constraints, not based on assumptions about standard languages, but on empirical documentation of dialectal or community usage, (c) determing whether code-switching as such is a functional mode of communication within the community, or simply an occasional artifact of interference or other language contact processes, and (d) assessing individual performance in terms of degree of community membership, degree of L_2 acquisition, and control of the code-switching mode.

Aside from the question of the validity of the two constraints across different multilingual communities, there is also the problem of additional constraints which might hold in specific social contexts. For example, in some situations involving clearly socially dominant/subordinate pairs of languages, switches may occur only by the insertion of occasional lexical items from the dominant language into the discourse of the other, but not the reverse (e.g. Denison 1972, G. Sankoff 1972). In the Puerto Rican situation the free morpheme constraint is partially superseded by a stronger constraint completely excluding English inflections on lexical items of Spanish origin, since such items rarely seem to be phonologically or semantically integrated into the English grammatical system (Pedraza ms.).

Another example from the Puerto Rican study involves code-switching among certain speakers whose migrational and educational history has resulted in their being less fluent in English than in Spanish. The equivalence constraint plays little role in this situation; because of their limited competence in English syntactic patterns, these speakers produce virtually no intra-sentential code-switches. Instead, they largely confine themselves to switching to English for sentence tags, interjections, and the occasional single noun in an otherwise entirely Spanish sentence (Poplack 1979a).

Indeed, the validity of any code-switching constraint, including the free morpheme and equivalence constraints, depends strongly on the particular configuration of social factors obtaining in a given community. A typology of the different patterns of code-switching would have to take account of such factors.

3. Scope of this study

Compared to the extensive literature on the interactional and pragmatic aspects of code-switching, the syntactic aspects have only begun to be clarified. One of the problems has been that the syntax involved is not easily or convincingly accessible to intuition; switches are not readily elicited, and acceptability judgements may be unreliable and normatively biased. On the other hand, observation is exceedingly difficult, given the precarious balance of

situational factors which must be sustained in order to assure the considerable volume of speech in the code-switching mode necessary for any statistically valid analysis of syntactic patterns.

One of the situational factors which may play a crucial role is the ethnic identity of the interviewer. As part of a long-term participant observation study in East Harlem, Pedro Pedraza collected recordings of Puerto Rican speech behavior in a variety of settings (Pedraza ms.). It has been demonstrated (Poplack 1978) that the in-group status of the interviewer coupled with relatively unobtrusive data gathering techniques yielded a body of code-switching data qualitatively more diverse and quantitatively more numerous than that which could have been elicited by an outsider to the community.

A selection was made of recordings of 20 individuals including both balanced bilinguals and speakers who are fluent in Spanish but not in English. The code-switches were extracted from these recordings with the help of Alicia Pousada, and were analyzed in a previous study (Poplack 1979a). An aim of this paper is to reanalyze these data within a formal grammatical framework. Because of the surface nature of the code-switching constraints described in section 1, the formalism we adopt is one based on the direct generation of surface phrase structures by a context-free grammar. In section 5 we justify our choice of this approach rather than an attempt to generate switches in deep structure. In order that this analysis be

as relevant as possible to the statistical generalizations drawm from speech performance data, we discuss how to probabilize the monolingual Spanish and English grammars, and the code-switching grammar which results from their combination. In a preliminary exercise based on speech samples of a Puerto Rican bilingual speaker, we then calculate the frequencies of the different rules in these grammars as well as the relative frequency of the various syntactic boundaries eligible to be the site of a code-switch. These frequencies of potential switch sites are then compared with actual switch frequencies at these sites compiled in the previous study for the sample of 20 Puerto Rican speakers, to give the relative susceptibility to code-switching of each kind of syntactic boundary.

The theoretical discussions serve as a framework and justification for this analysis of syntactic boundaries and their switch propensities, which is the main innovation of this paper. For the first time, we present actual code-switching rates, and these show that the equivalence and free morpheme constraints have implications which go beyond their qualitative formulations.

4. How many grammars?

There has been some debate over whether discourse containing code-switches is generated by the alternate use of the two monolingual grammars or whether a single code-switching grammar

exists, combining elements of the monolingual grammars.[2] There are really two questions involved, one notational or definitional, and one substantive. Any finite set of rules and procedures for generating an infinite set is a grammar, formally speaking, so that any set of rules for constructing the set of sentences containing code-switches is a grammar.[3]

Apart from definitions of a grammar, there remains the more important question of whether code-switching involves the alternation from one distinct linguistic system to another, or whether speakers are exemplifying some integrated competence in the two languages. The evidence which seems most pertinent to this issue is the finding that code-switching generally does not entail pauses, hesitations, repetitions, corrections or any other interruption or disruption in the rhythm of speech (Poplack 1979a). This is distinct from many bilingual situations marked by language interference, for example, and provides some justification for treating code-switched discourse, at least in parts, as being generated by a single grammar based on the two monolingual ones. It will be clear, moreover, from the way that this grammar must be constructed, that code-switching is not a result of imperfect competence in either of the two monolingual modes of communication but rather results from knowledge of the rules of both, their similarities and differences; nor do code-switchers suffer loss of competence by virtue of their skill at the code-switching mode.

One way of avoiding the conceptual problems involved in the notion of a code-switching grammar has been to postulate that one or the other of the monolingual grammars is basic to any particular sentence. But this attitude, embodied in the hypothesis advanced by Wentz (1977) and others that every sentence has only one "base" language, which can be ascertained by the languages of the determiner and/or the verb, does not seem pertinent to the East Harlem situation (nor, for that matter, to other published Chicano data). The viewpoint that there is an easily identifiable base language is associated with the notion that code-switches involve the insertion of isolated L_1 elements or constituents in otherwise L_2 discourse, or vice-versa. This may very well be the case in certain contexts, such as those described in some of the studies cited in section 2. Indeed, in the previous analysis of the Puerto Rican data, a method was operationalized to identify "base language" and "language of the switch". It became clear, however, that in many cases this procedure was arbitrary.

A sketch of the different types of distribution of the two languages in code-switching discourse will help explain why. Such discourse may contain a stretch of several sentences clearly identifiable as belonging to one language (except for occasional words or constituents), as in (5).

(5) 'Cause I believe they're poor, they gotta know how to eat everything; not just little desserts and *esos potes* [those jars] which I don't like them. (S.L./9)

But in other stretches, constituents may oscillate several times from one language to the other, even within the confines of a single sentence, as in (6).

(6) There was a guy, you know, *que* [that] he *se montó* [got up]. He started playing with *congas*, you know, and *se montó y empezó a brincar* [got up and started to jump] and all this shit. (P.R./25)

There is no empirical justification for insisting that stretches like (6) or (17) have one underlying language with insertions from the other language. Indeed, no algorithm to determine "base language" so far proposed applies consistently and convincingly to performance data containing multiply switched sentences. What is more consistent with the data is simply to allow the possibility that in the uttering of a sentence, the rules used to construct its constituents may be drawn at times from one monolingual grammar and at times from another. Thus in what follows, neither the root S node of a phrase structure tree, nor the NP, VP, etc. nodes, must be identified as to language, though some of them necessarily will be.

Summarizing these considerations, long monolingual stretches of discourse may be thought of as being generated by a monolingual grammar, but the notion of a code-switching grammar seems to be called for where switches occur with high density. It will be seen that such a grammar may be formalized so as to subsume the two monolingual grammars, allowing the entire discourse to be analyzed in a uniform framework.

5. Switches--surface or deep?

The code-switching constraints are constraints on the surface syntax of a sentence. There is no empirical evidence that code-switched sentences are generated as such in a base component and preserved as such through a series of transformations, as suggested, for example, by Barkin and Rivas (1979). Indeed, the evidence is against this. Parts of sentences which may be analyzed as having been displaced by movement transformations are in no way constrained, in real data, to be of the same language as the elements which may have been adjacent to them in deep structure, but are rather constrained, if at all, by their surface neighbors.

The following example is somewhat of a straw man, since both of its postulates are easily demolished. However, it clearly illustrates how a movement transformation operating on a constituent which is constrained against being switched in deep structure, implies a clearly invalid surface constraint.

Timm (1975, 1978), Gumperz (1976) and Barkin and Rivas (1979) have all suggested that underlying subject pronouns must be in the same language as the verb of a sentence. Thus the code-switch in sentence (7) below (as well as one in (6)) should be excluded. Were passives generated transformationally, sentence (8) would also be excluded since its underlying form is of the same type as (7). In fact, (8) is not excluded, being typical of attested code-switches involving prepositional phrases.

(7) You *estás diciéndole la pregunta* in the wrong person.
[You're asking the question to the wrong person.]
(P.A./43)

(8) La pregunta fue dicha [the question was asked] *by you.*

The facts that sentences like (7) are also attested in these data, and that passives are not transformationally generated in many current analyses, do not alter our contention that a transformational analysis of code-switching will necessarily exclude many well-attested constructions.[4] Conversely, such an analysis might also produce violations of the code-switching constraints by moving items remote in deep structure, and hence permitted to be in different languages, to adjacent positions on the surface, where they would violate the free morpheme or equivalence constraints:

(9) The car *del hombre* [of the man].

but

(10) *el hombre'*s car*

6. Free union grammar

Following the considerations of the preceding sections, we will seek in the data analysis to sketch surface grammars for the Spanish and English spoken in our corpus, as well as for the code-switching mode. Our goal is obviously not to solve all the classical problems involved in constructing a complete generative description of any of the languages involved, but to illustrate how two formal monolingual grammars can be combined to produce a grammar of the code-switching mode.

Suppose we have two context-free phrase structure grammars G_1 and G_2 for two languages L_1 and L_2, such that the non-terminal grammatical categories of one generally have corresponding categories in the other. We call this the first translatability condition. In addition, we assume each rule in G_1 can be functionally translated by at least one rule in G_2, e.g. the rule S → VP NP which results in Spanish post-posed subjects can always be translated by the English S → NP VP. This is the second translatability condition. These two translatability conditions will generally hold for any two natural languages described within a common theoretical framework.

The first condition allows us to define the free union of the two grammars consisting of the common set of grammatical categories, the combined set of rewrite rules from G_1 and G_2, and the combined lexicons. The resulting entity is a phrase structure grammar, it is context-free, it subsumes the two monolingual grammars, generating all sentences in L_1 and L_2, and it generates all possible sentences containing code-switches. Yet this grammar is of little interest. Not only does it generate equivalence constraint violations like (11), but it also generates ungrammatical monolingual constituents like (12).

(11) NP → DET N ADJ (from Spanish)

DET → the
N → casa
ADJ → white

* the *casa* white

(12) NP → DET N ADJ (from Spanish)

DET → the
N → house
ADJ → white

* the house white

Thus the free union grammar of G_1 and G_2 is not a satisfactory code-switching grammar. Some mechanism must be found for restricting the output of the grammar so that the monolingual sentences it generates are grammatical according to G_1 or G_2, and the bilingual sentences satisfy the code-switching constraints. One way of doing this would be simply to have an output filter which rejected all unsuitable sentences. In general, however, the problem of constructing a finite set of rules for recognizing ungrammatical sentences, or switches violating the constraints, is no less difficult than constructing the entire grammar. This solution, then, would only be feasible for some special pairs of very similar languages where code-switching violations could be easily recognizable as belonging to some small predetermined set. Furthermore, this solution not only trivializes the problem of finding the structure of the code-switching grammar, but also results in a grammar which is not context-free. Rather, it has some ill-defined, complicated structure which is not directly comparable to the monolingual grammars.

Having thus rejected the free union grammar, with or without output constraints, we are faced with the key task of this paper: to incorporate the code-switching constraints into the rules of the phrase-structure grammar without altering its context-free nature.

The basic problem is that the code-switching constraints are, generally speaking, conditions on adjacent constituents, but the essence of context-free generation of sentences is that the internal structure of one constituent does not condition that of another. To solve this problem we must ensure that for any two neighboring constituents whose boundary could potentially involve a code-switch violating one of the constraints, suitable restrictions must already be coded into the symbol for the grammatical category heading each constituent. And as these symbols are rewritten, the restriction information must be passed on to, or inherited by, lower level constituents, so that when the terminal grammatical categories are finally lexicalized, the restriction will be realized by a compatible choice of language for neighboring lexical terms.

The approach we will take is to introduce superscripts on the symbols for the various categories of the grammar, and to restrict the application of certain rules to symbols with appropriate superscripts. These superscripts will appear only in certain derivations and only at certain nodes, and they will carry information sufficient to prevent any violation of the code-switching constraints, and to permit any code-switches which do not violate them.

7. A code-switching grammar

The code-switching grammar will then be constructed as follows. Its lexicon will be the combined lexicon of the two monolingual grammars. Its grammatical categories will be the grammatical categories of G_1 and G_2 (most of which they have in common). Each category may occur in a (possibly large, but finite) number of versions, depending on the presence of superscripts, as will be explained below. As for the rules of the code-switching grammar, consider first any rule R in G_1. Using the second translatability condition stated above, we can compare R to all its possible translations by rules of G_2. Suppose for any pair of symbols in the output of R, there exists at least one G_2 translation which does not reverse the order of the two symbols. Then R is included among the rules of the code-switching grammar, again possibly in a number of different versions. Rules of G_2 are similarly included in the code-switching grammar if they satisfy an analogous condition. Now, if in the output of the rule R there are two (obligatory) symbols ordered in a way excluded in all the corresponding G_2 rules, the equivalence constraint means that we must not allow a switch from L_1 to L_2 after the constituent headed by the first of these two symbols, and before a constituent headed by the second. The free morpheme constraint, on the other hand, means that if in the output of the G_1 rule, there are two consecutive symbols, the first of which represents a morpheme bound to the second, a switch from L_1 to

L_2 must be precluded between the symbols. Likewise, if the second symbol is the bound morpheme, no switch from L_2 to L_1 may intervene.

To ensure that such restrictions are obeyed, the rule R cannot be incorporated into the code-switching grammar as is. Rather, the two symbols in question must be modified in the output of R by superscripts which indicate that the constituents they head are in a strictly L_1 order.

8. Superscript conventions

Each superscript will have two components separated by a colon, the first component indicating a language, the second a terminal category (e.g. sp:adj or eng:det). This category and only this category will be the one which must be lexicalized in L_1. A simple example involves the Spanish rule NP → DET N ADJ, whose English translation is NP → DET ADJ N (13). Here the superscript on the N in the code-switching grammar is $^{sp:n}$, and on the ADJ it is $^{sp:adj}$. In this case the superscript means only that when the category N is lexicalized, it will be in Spanish, and similarly for the ADJ. Note that the DET remains unsuperscripted, so that it may be lexicalized in Spanish or in English.

(13) Spanish: NP → DET N ADJ

English: NP → DET ADJ N

Code-switching: NP → DET $N^{sp:n}$ $ADJ^{sp:adj}$

This suffices to preclude code-switching constraint violations like (11) and monolingual grammaticality violations like (12).

To satisfy the free morpheme constraint, it is necessary that any rule generating a Spanish bound morpheme incorporate "sp" superscripts on this morpheme and on the free morpheme category to which it is bound.

What of rules rewriting higher order categories? For the Spanish postposed subject rule in (13a) we cannot leave unrestricted further choice of rewrite and lexicalization rules without risking generating sentences like (13b). Nor do we want to be restricted to Spanish only for further rules: this would exclude (13c) which is in no way unusual.

(13a) S → VP NP

(13b) *arrived he

(13c) Llegó *yesterday* la mamá mía. [My mother arrived yesterday.]

Thus the grammatical category component of the superscript must be carefully chosen to ensure that the equivalence constraint is not violated, but without putting any other restriction on the string being generated. This is done as in (14).

(14) $S \rightarrow VP^{sp:v}\ NP^{sp:l}$

When the VP is rewritten, its superscript is transmitted to all symbols in the output of the rewrite rule, as in (15).

(15) $VP^{sp:v} \rightarrow V^{sp:v}\ ADV^{sp:v}$

When the $V^{sp:v}$ is lexicalised, it must be in Spanish, but as for the $ADV^{sp:v}$ category, since the superscript does not specify $^{sp:adv}$, an adverb may be chosen from either the Spanish or English lexicons--cf. (13c). We refer to this as a heritability condition. The transmission of the $^{sp:v}$ superscript from any symbol which has it to all the symbols which rewrite it is the most general type of heritability condition.[5] For each rule which rewrites VP, another version must occur in the grammar with all symbols superscripted $^{sp:v}$, and the same holds for any symbol in their outputs which is non-terminal (i.e. is to be rewritten), and so on. The only exceptions are: (a) embedded S nodes do not inherit superscripts, (b) superscripts originating in equivalence constraints in embedded constituents, or in free morpheme constraints, supersede those from higher order equivalence constraints and (c) lexicalization of categories not involved in the superscript is unrestricted as to language.

A second type of heritability condition is exemplified by the $^{sp:1}$ superscript on the NP in (14). Any time a symbol superscripted this way is rewritten, the superscript must be passed on to at least one symbol in the output of the rule. And any terminal grammatical category thus superscripted must be lexicalized in Spanish. Again, embedded S nodes do not inherit this superscript. The $^{sp:1}$ superscript serves simply to ensure that the NP is not entirely lexicalized in English--though there are no empirical

grounds for specifying that any particular element of the NP, even the DET, be in Spanish.

In another example describing Spanish conjoined noun phrases both modified by a shared adjective (16a), the rule must be re-specified as (16b), so that the CONJ and any element of each of the conjoined NPs, other than the N, may be switched to English. The $^{sp:n}$ superscript is of the same type as the $^{sp:v}$ superscript in (15) and has the same heritability condition.

(16a) Spanish conjoined NP: NP → NP CONJ NP ADJ

(16b) Code-switching: $NP \rightarrow NP^{sp:n}\ CONJ\ NP^{sp:n}\ ADJ^{sp:adj}$

Are any other types of superscripts involving different heritability conditions necessary? In the present study we have not found any necessary, but this may simply be a function of the two languages involved, and of the precise way the free morpheme and equivalence constraints function for particular pairs of languages. Thus our procedure for constructing the set of rules in the code-switching grammar may have to be modified as different types of non-equivalence are examined.[6] The fundamental principle, however, will remain the systematic comparison of corresponding G_1 and G_2 rules.

Every time a discrepancy between G_1 and G_2 higher order rules may lead to a violation of the code-switching constraints, we must first identify the constituents which risk being involved in this violation. We then incorporate lexicalization restrictions in the

terms of the higher order rule, restrictions which carry with them certain heritability conditions to ensure that lexicalization is carried out appropriately but is not overly constrained. This entails a proliferation of categories and rules in the grammar, but does not interfere with its context-free nature. Note that the restrictions are a function of the similarities and differences between the two languages involved, and derive only from the equivalence and free morpheme constraints and not from any other purported universal syntactic properties of VPs, for example.

9. Probabilistic grammars

In the remainder of this paper, we will analyze the syntactic aspects of code-switching heard in the speech of Puerto Rican bilinguals. Though the context-free grammar for the code-switching mode described above may well account for the types of switches allowed and those excluded in this corpus, it cannot by itself capture many of the other regularities observed in this type of discourse. In particular, and it shares this inability with any generative grammar when confronted with performance data, it cannot account for the many striking quantitative patterns evident in the discourse.

A grammar will, however, generate the quantitative structure of a language as well as its qualitative or categorical aspects, if a suitable probabilistic component is added to the generative

machinery. Context-free grammars are easily probabilized, as noted years ago by e.g. Klein (1965) and Grenander (1967). Probabilistic context-free grammars have been used to study style-shifting (Klein 1965), first language acquisition (Suppes 1970), grammatical inference (Horning 1969, Sankoff 1971, 1972), the acquisition of German by migrant workers (Heidelberger Forschungsprojekt "Pidgin-Deutsch" 1978, Klein and Dittmar 1979), and differences in noun phrase structure in written and spoken English (Hindle 1980).

In this section we will discuss the relationship between the probabilistic context-free grammar generating L_1 and L_2 monolingual speech on the one hand, and the probabilistic code-switching grammar on the other. This will serve as a conceptual framework for the analysis in the next sections.

The key to the probabilization of a context-free grammar is that when a node of a given category is to be rewritten, the choice of rewrite rule is made according to a set of probabilities over all possible rules for rewriting that category, and is made independent of all other choices of rewrite rules in the derivation. Thus if the only ways to rewrite NP in a grammar were summarized by

$$NP \rightarrow (DET)\ N\ (ADJ),$$

then each of the possibilities NP → N, NP → DET N, NP → N ADJ and NP → DET N ADJ would be assigned a probability, i.e. a number between zero and one, in the definition of the grammar, and these

numbers would have to sum to one. Then every time an NP was to be rewritten, a random (not to be confused with equiprobable) choice among the four possibilities would be made with each one's chances of being chosen equal to its associated probability. A similar set of probabilities would exist for the rules rewriting S, another set for VP, and so on.

For a given context-free grammar the rule probabilities can be estimated by examining a sufficiently large corpus, or sample of the generated language, parsing each sentence, and counting rule frequencies. If there are ambiguities, more complicated procedures are necessary (Sankoff 1971, 1972).

Our conditions in the previous section on the translatability of categories of G_1 and G_2 mean that they are essentially two probabilistic context-free grammars using the same set of symbols, and this led to a natural definition of the code-switching grammar. Complications arise when we come to probabilize the rules of this new grammar. How are the probabilities associated with the rules of G_1 and G_2 combined to produce the probabilities of the rules in the code-switching grammar? To answer this question will require a great deal of empirical research. Once sufficient data analysis enables us to establish the mechanism for combining probabilities, this mechanism will be the key to truly integrated deductive/inductive research on the relationship between probabilistic monolingual and code-switching grammars. That is, the statistical

properties of the code-switching grammar will not only be empirically observable in code-switching discourse, but will also be predictable from the statistical properties of the monolingual grammars. The format of the data available to us, however, and the preliminary nature of this exercise, permit us access to code-switching statistics only by directly examining code-switching discourse, and not by deduction from the monolingual grammars. For the present we can only speculate on the details of the probabilistic mechanisms involved in combining grammars.

The simplest hypothesis takes account of the observation that a given stretch of code-switching discourse is characterized by a certain proportion of L_1 and a certain proportion of L_2. These proportions are sensitive, among other things, to the bilingual ability of the speaker, and the nature of the interlocutor, situation and topic, but even with all such factors held constant, basically monolingual stretches alternate with stretches of high code-switching density, as mentioned in section 4.

The hypothesis would have rules for rewriting a category in the code-switching grammar chosen at random from the eligible rules in G_1 and G_2, with the probabilities being a compromise between the probabilities in the two monolingual grammars, weighted according to the proportion of L_1 and L_2 in the overall discourse. (There would be exceptions, of course, especially when certain superscripted categories were rewritten.) It seems likely, however,

though this would need to be verified mathematically and experimentally, that this choice mechanism would yield far more multiply switched sentences than are empirically observed. To circumvent this difficulty, it will probably be necessary to allow ^{sp} or ^{eng} superscripts on some phrase structure nodes aside from those discussed in section 8. When a node is to be rewritten, each subcategory will be superscripted in the same way (or each lexicalization would be in the same language). Depending on the prevalence of such superscripted nodes, we can obtain rates of code-switching per sentence more in accord with observed tendencies. The complete solution of this problem awaits further quantitative research, but later we will present empirical evidence that the code-switching grammar probabilities do represent compromises between the two monolingual grammars.

Why are the rule probabilities of the code-switching grammar so important? It is because the probabilities in a context-free grammar determine <u>all</u> the statistical and quantitative properties of the language it generates. In particular, they completely determine the preferred locations and frequency of code-switches within the sentence. And it is the comparison of these theoretical predictions with the type of observations and calculations in the next section which is the most promising way of verifying a formal syntax of code-switching.

10. Code-switching frequencies and rates

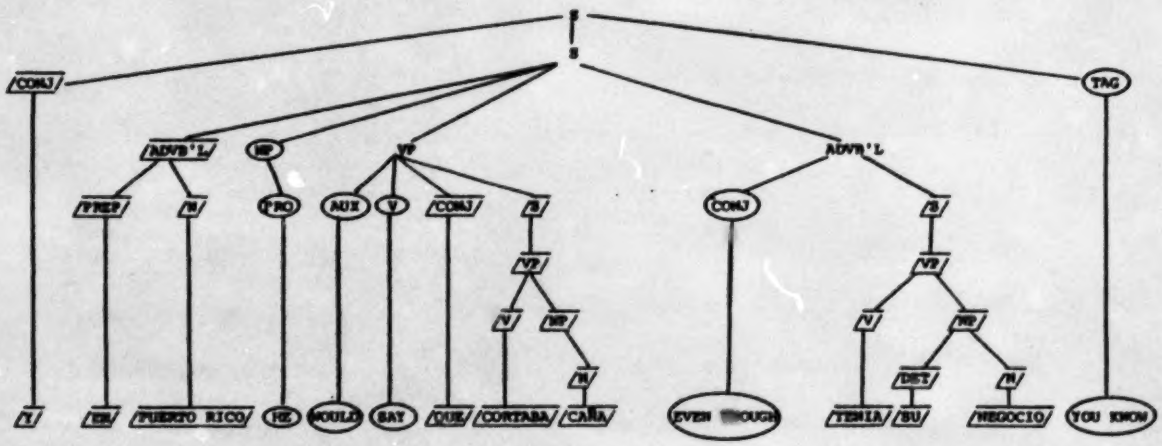
In the study of code-switching it does not suffice to document the rarity of exceptions to purported syntactic constraints in order to prove them. For example, Timm (1978) attempted to validate the universality of the syntactic constraints she earlier (1975) felt to be valid for Spanish-English switching, by counting the exceptions to these constraints in Russian-French code-switching discourse in Tolstoy's War and Peace. For most of the constraints conjectured she found only a few exceptions. However, since she does not indicate how much code-switching discourse is contained in the opus or how many code-switches there are in all, or how many are intra-sentential, the significance of the exceptions cannot be assessed.

Previous quantitative studies (Pfaff 1975, 1976; Poplack 1978, 1979a) have been more revealing in showing what proportion of code-switches involved nouns, what proportion determiners, etc. Even this, however, does not give a clear indication of the quantitative effects of syntactic context on code-switching. Just because single nouns, for example, were found to constitute 14% of the switches, while predicate adjectives made up only 3%, this does not necessarily mean that nouns are more likely to be switched than predicate adjectives. It could simply be that nouns occur five or ten times more frequently in discourse than predicate adjectives. To estimate the true relative susceptibility of a syntactic boundary to be

the site of a code-switch, then, we must divide the raw frequency of switches at each type of boundary by the frequency of occurrence of this boundary in the code-switcher's discourse.

Thus we undertook to estimate the overall occurrence rate of various constituent boundary types in typical discourse containing code-switches. To do this we isolated, in a series of recorded conversations with a balanced bilingual speaker, some 30 stretches of discourse containing code-switches. The one or more sentences in each stretch were parsed using a limited number of syntactic categories, as in (17).

(17) Y en Puerto Rico *he would say* que cortaba caña, *even though* tenía su negocio, *you know*. [And in Puerto Rico he would say that he cut cane even though he had his own business, you know.] (S.L./32)



Key: ▱ : Spanish

◯ : English

In accordance with the discussion at the end of section 9 above, we also attempted to infer which nodes of the phrase structure tree could be unequivocally identified as to language. The following criterion was adopted: whenever a node dominated only Spanish lexical terms, the rule rewriting it was clasified as a Spanish rule, and analogously for English rules. The remainder, those that dominated both Spanish and English (in example (17), the S, VP and ADVB'L nodes) lexical items, were listed separately as most representative of the code-switching mode.

This manner of identifying node languages applies more widely than the language choices required by the code-switching constraints discussed in sections 7 and 8. In example (17), the only superscripts imposed by the equivalence constraint would be $^{sp:v}$ on all nodes of the VPs *cortaba caña* and *tenía su negocio*, reflecting the impossibility of subject pronoun absence in English in this context. As argued in section 9, the additional specification of language in some nodes not involved in code-switching constraints may help to better account for observed rates of code-switching. Although in parsing we can easily identify the nodes, further research will be necessary before we can suggest a probabilistic mechanism for the choice of such nodes in the generation process.

11. Rule frequencies

Using the surface phrase structures obtained from the parsing procedure, we were able to tabulate (a) the frequency of the various rewrite rules used in generating the sentences, and (b) the frequency of constituent boundaries of various types.

As in the previous section, we point out that theoretically, we would want to use the estimates derived from the Spanish only and English only data to predict rule probabilities in the code-switching mode. Further, we would like to predict the frequency of various constituent boundaries as well as the switch frequency at these boundaries. These predictions could then be compared to the empirical results with a view to further refining the theory.

As discussed in the previous section, however, neither our knowledge of the generative machinery, nor these preliminary data, are sufficient for detailed inference based on a probabilistic context-free grammar model. In Table 1, however, we can make some inter-code distinctions by separating rewrite rules applying to nodes identified as English and those identified as Spanish. The remainder are listed under 'code-switching'. Certain differences are obvious in the Table.

	English	Spanish	Code-switching
S → {TAG / ADVB'L / Ø} NP VP {TAG / ADVB'L / Ø}	79%	20%	62%
S → (TAG) VP (TAG)	6	80	15
S → (CONJ) NP	3	0	3
S → VP NP (TAG)	0	0	6
S → CONJ S {TAG / ADVB'L / Ø}	12	0	9
S → NP S (TAG)	0	0	6
	(n = 33)	(n = 10)	(n = 34)
NP → PRO	60	9	0
NP → (DET) N	20	70	43
NP → (DET) ADJ* N	9	6	7
NP → (DET) N ADJ'L*	3	15	0
NP → NP S	3	0	29
NP → NP CONJ NP	5	0	21
	(n = 68)	(n = 33)	(n = 14)
VP → (MOD) V NP (ADVB'L)	31	35	21
VP → (ADV) (MOD) V {PREP PHR / ADVB'L / ADJ'L / Ø}	38	18	14
VP → AUX (NEG) V {NP / SUB CONJ S / PREP PHR / S / Ø}	31	0	14
VP → PRO*+ (MOD) V (NP)	0	30	0
VP → NP+ V	0	11	0
VP → NEG V NP	0	6	7
VP → (AUX) V {S / SUB CONJ S}	0	0	43
	(n = 32)	(n = 17)	(n = 14)

Table 1. Probabilistic phrase structure grammars. S, NP and VP rewrite rules for English, Spanish and code-switching modes.
* One or more constituents of this type.
+ These are object NPs, pronouns, or reflexive clitics.

The most striking distinction in Table 1 is the tendency for English sentences to be derived by a S→NP VP rule followed by NP→PRO, whereas in Spanish the dominant tendency is for S→VP, followed by NP→(DET) N. This difference reflects the prevalent option for subject pronoun deletion in Spanish.

In the NP rules we note a difference between adjective or adjectival placement in Spanish and English. This is only a quantitative difference--though most Spanish adjectives must be postposed, some may precede the noun, and both English and Spanish adjectivals follow the noun.

In the VP rules, we note the difference between Spanish and English in the use of verb auxiliaries and negation, and in the position of the object NP. In Spanish but not English, the object NP is optionally preposed, and obligatorily preposed in many cases when it is pronominalized.

There is a general tendency for the numbers in the code-switching column to resemble the English figures in some respects and the Spanish in others. The exceptions result from two factors. One is simply statistical fluctuation due to the sample size. The other, more important, is the apparent elevated tendency for recursive rules involving subordination and conjunction to be employed in the code-switching mode when rewriting S, NP and VP. This latter tendency is probably largely due to the fact that those rules used late in the derivation, containing few embedded

constituents, were most likely to be clearly monolingual, i.e. Spanish or English, while those rules used earlier, generating conjoined and subjoined structures, dominated many more constituents and were thus more likely to dominate constituents of both languages, so that they could not be inferred to be drawn from either the Spanish or the English grammar.

An important conclusion to be drawn from this part of the exercise is that even in those portions of discourse in close proximity to one or more code-switches, the speaker is strictly maintaining the qualitative and quantitative distinctions between the Spanish and English grammars. Whenever a stretch of discourse, no matter how short, can be clearly identified as monolingual, the rules of the appropriate monolingual grammar, and their associated probabilities, are exclusively in play.

12. Code-switching rates

Along with the rewrite rules discussed in the previous section, the parsing exercise also produced frequency counts of constituent boundaries of various types in the 30 discourse stretches analyzed. In this section, we combine these data with the results of Poplack (1979a) on the observed frequencies of switches of various grammatical categories, in order to evaluate switch rates, i.e. the propensity of given syntactic boundaries to be the site of a code-switch.

Had the latter data been in terms of switch frequencies at the various constituent boundaries, and had the two data sets been compiled on exactly the same corpus, it would have been an easy matter to divide the switch frequency at each boundary type by the frequency of occurrence of that boundary type, and hence, to derive the switch rate for that type of constituent boundary.

But because the 1979a data were compiled in terms of the grammatical category of the switched item, we first had to convert them to boundary terms by cross-tabulating the category of the switched item with the categories of the preceding and following items.

And because the corpus for the category frequency data was not identical to the corpus for the boundary frequency data, dividing the former (converted from category to boundary terms) by the latter does not give the switch rate, but a number which must be multiplied by a certain factor to obtain the switch rate. This factor is largely determined by the relative size of the two corpora, and remains constant for all boundary types, since the same disproportion between the two corpora holds for the data from each type of syntactic boundary.

This means that even if the numbers obtained by dividing switch frequencies by boundary frequencies are not the actual switch rates, they are all proportional to the 'true' switch rates by the same constant of proportionality.

In any case, we have already noted that in code-switching discourse, rates are by no means homogeneous, either from situation to situation, or from speaker to speaker. Thus, dividing switch frequencies by boundary frequencies for the whole corpus, including the largely monolingual parts, would have produced rates too low for stretches where switches are dense, and too high for stretches where they are rare. (Again, however, 'too high' and 'too low' would apply uniformly across all boundary types, so that if the estimated rates are not really applicable to a given stretch of discourse, they are all proportional to the true rates.)

Moreover, given that a speaker's propensity for switching differs according to both extralinguistic factors and the specifics of the given conversational interaction, the calculation of absolute, or universal, switch rates does not seem to be a very meaningful goal. But since we cannot expect any interaction between these extra-linguistic factors and the boundary types affected by switching (with one exception, to be discussed below), changing the situation will change the switch rates, but only in a proportional way across all boundary types.

In sum, our primary goal must be to calculate not the switch rates themselves, but the ratios between the switch rates at various syntactic boundaries. As the situation changes, or the speaker changes, or even from one stretch of conversation to another, the switch rates will all change, but will remain in the same proportion

to each other. Thus we need not be overly concerned about the fact that our calculations only produce figures proportional to code-switching rates rather than the rates themselves, since it is only the proportionality among the rates which can hold throughout a discourse, from speaker to speaker and from situation to situation.

Thus in Table 2, we show the relative propensity for each syntactic boundary type to be the site of a switch, using the formula in (18).

$$(18)\quad \text{code-switch rate at a given syntactic boundary} = \text{constant} \times \frac{\text{number of switches at boundary}}{\text{frequency of boundary}}$$

For Table 2, a constant was chosen in an effort to obtain the probability of a code-switch at a given syntactic boundary in a typical stretch of code-switching discourse. The figures in Table 2 are, as we have stressed, meaningful in a proportional sense only, i.e. they may all be too high or too low by a constant factor, and this factor will change from situation to situation, and from speaker to speaker.

SWITCH SITE	RATES
Between tag and preceding or following category	40 %
Between ADV and ADVB'L and preceding or following category	5-10
Between PRED ADJ and preceding category	15
Between DET and N or NP	13
Between coordinate conjunction and preceding category	9
Between subordinate conjunction and following category	3.9
Between $\{^{V}_{VP}\}$ and $\{^{N}_{NP}\}$	2.7-3.6
Between coordinate conjunction and following category	2.2
Between ADJ and $\{^{N}_{NP}\}$	2.2
Between $\{^{N}_{NP}\}$ and $\{^{V}_{VP}\}$	2.3
Between PREP and following category	2.3
Between $\{^{V}_{VP}\}$ and PREP PHRASE	2.3
Between $\{^{AUX}_{MOD}\}$ and $\{^{V}_{VP}\}$	.9
Between PREP PHRASE and ADJ'L (except after $\{^{V}_{VP}\}$ and preceding category)	<.1
Between subordinate conjunction and preceding category	<.2
Between pronoun and preceding or following category	<.1
Between clitics and V	0
Between AUX and NEG	0
Between NEG and $\{^{V}_{MOD}\}$	0
Between VP and subject NP	0

Table 2. Code-switching rates at different syntactic boundaries.

Table 2 shows that constituent boundaries are clearly subject to a hierarchy ranging from very high propensity to be the site of a switch, to total absence of switching. We remark first that prohibited switch sites are precisely those in the vicinity of which the number and/or order of sentence elements generated by a given rule is excluded in one of the two languages, i.e. those which

violate the equivalence constraint. Included here are constructions involving NEG placement, which in Spanish directly precedes the main verb, as in (19), while in English it follows an auxiliary or a modal, as in (20).

(19) An' the second one, I seen everything 'cause *no cogí na'* [I didn't take anything]. (S.L./1)

(20) La anestesia [the anesthesia], *I didn't take it.* (S.L./2)

Also included here are constructions involving reflexive and object pronoun clitic placement, which in Spanish precede the verb, as in (21), and in English follow if they appear in the surface structure at all; similarly, for Spanish constructions in which the subject NP follows the verb, as in (22). Switches in these examples may occur around, but not at, the boundaries in question.

(21) This one, he doesn't wanna eat *casi*, right? *Se le da un dolor de barriga* [he gets a stomach ache]. He gets a lot of stomach pains. (S.L./4)

(22) I really been in here, which *quería Juan* [Juan wanted] you know, *desde* [since] nineteen seventy two. (S.L./28, 29)

Note that these restrictions stem only from the differences between two languages involved in the code-switching mode. Published data on French-Italian, for example, which both make use of equivalent rules of clitic pronoun placement, include a switch between clitic and verb: *si* sent 'S/he feels' (di Sciullo et al. 1976).

At the other extreme, the greatest propensity to switch is

shown by the category TAG, both when it precedes and follows each of the 16 other syntactic categories studied, as in (23), for example, and despite the fact that this segment occurs relatively rarely in non-code-switched discourse.

(23) Yo estaba aburrecido, muriéndome, *you know*? [I was dying of boredom, you know?] (C.B./28)

This reflects the fact that tags are subject to minimal if any syntactic restrictions and so may be switched easily without fear of violating the equivalence constraint. Indeed, switches of precisely this category were found (Poplack 1979a) to characterize the discourse of non-fluent bilinguals, allowing them to participate in the code-switching mode although they lacked the bilingual ability in L_2 to engage in more complex switching.

If any boundary types do not obey the proportionality relationship discussed above, it will be those involving tags. Thus for certain speakers, switches involving tags will be increased dramatically, while those involving other constituents will not only not increase proportionally, but may even decrease. The 40% figure attached to tags may be somewhat exaggerated relative to the other rates because the 'category frequency' data on which they are based contained many tags switched by non-fluent bilinguals who engaged in little other intra-sentential code-switching.

Another favored switch point is before a predicate adjective. This preference contrasts sharply with the restrictions against

switching between the non-equivalent noun + adjective or adjective + noun boundaries to be discussed below.

The point between determiner and noun will be the site of a switch about 13% of the time according to these calculations, a finding reflecting the great susceptibility of nouns not only to be borrowed, but also to be switched, as is widely noted in the literature (e.g. Weinreich 1953, Gumperz 1971, Timm 1975, Wentz 1977).

Finally, adverbs and adverbial phrases, both preceding and following the other constituents studied, are very likely to be switched, with a rate of 5-10% depending on the specific constituent with which they are combined. This again reflects, though not as strikingly as for tags, the large number of slots these categories may occupy (as in in (24), for example) within the sentence without fear of violating the equivalence constraint.

(24a) A los cuatro meses [at four months] *they start munching on some rice and beans.* (S.L./8)

(24b) Uno no podía comer carne [we couldn't eat meat] *every day.* (S.L./20)

Conjunctions and prepositions show an interesting pattern of asymmetries in these data. Coordinate conjunctions tend to be in the language of the following constituent, as evidenced by the high propensity to switch before such constituents in contrast with an average propensity to switch after them. Subordinate conjunctions and prepositions, however, tend strongly to remain in the language

of the head element on which they depend, and it is the remainder of the dependent clause which is switched. This switch rate would seem to tie in with Gumperz' (1976) constraint requiring that conjunctions be in the same code as the conjoined sentence, at least insofar as coordinate conjunctions are concerned. Why coordinate and subordinate conjunctions should behave distinctly in this regard, however, is not immediately apparent. Nor is the data conclusive. Examples such as (25) are not rare. For the moment, then, we must allow for the possibility that the quantitative patterns are due to sparse data.

(25a) I could understand *que* [that] you don't know how to speak Spanish, *¿verdad?* [right]? (S.L./75)

(25b) Right to 104th Street *donde tenía una casa* [where I had a house] which were furnished rooms. (S.L./25)

(25c) Any kind of book that's interesting, about Mafia *o* [or] love story *o* sex books or things like that. (W.B./27)

If the tendency of switches to occur after and not before prepositions and subordinate conjunctions is borne out, however, this would dispel any identification of higher order constituent boundaries with ease of switching, and constituents linked by late rewrite rules with resistance to switching: prepositions and subordinate conjunctions are both linked at a higher level with their header categories than with what follows them.

The boundary between verb and following object NP shows a somewhat higher switch rate than that between preceding subject NP and following VP, though both types of switches are far more frequent

than any before or after a subject pronoun. Indeed, it is precisely the very low propensity of subject pronouns to be switched which explains why scholars have posited categorical constraints against switching them (e.g. Timm 1975, 1978; Gumperz 1976), and which most clearly illustrates the utility of a quantitative approach to the study of code-switching.

We remark that a large proportion of syntactic boundaries are affected by the same, intermediate switch rate of approximately 2.2 - 2.3%. Now, if the equivalence and bound morpheme constraints were not only qualitatively but also quantitatively the only constraints on code-switching, we would expect all switch rates for all boundaries to be the same. And indeed, apart from the especially susceptible types, largely involving freely moveable constituents, and the very low frequency types, which in some cases seem to approach being morphemic rather than syntactic boundaries, all other constituent boundaries involve switches at a rate proportional to the frequency of these boundaries in monolingual speech.

We find that even the boundary between adjective and noun has an intermediate switch rate, i.e. the propensity for this boundary to be the site of a switch largely reflects its frequency of occurrence in non-switched discourse. This is somewhat surprising since most Spanish adjectives do not follow the equivalent word order, as may be seen in (26).

(26a) No coge la estación latina. [It doesn't get the Latin station.] (W.B./23)

(26b) Because they're Spanish people. (W.B./62)

Many do, however, and at any rate, this switch site has already been shown (Poplack 1979a) to represent the majority of the few attested violations of the equivalence constraint.

Showing a relatively low propensity to be the site of a switch is the point between auxiliary or modal and verb, which again explains why categorical constraints have been posited (Timm 1975, 1978) against switching here.

13. Discussion

In constructing a formal apparatus as a framework for the empirical exercise, several points emerge. The code-switching constraints are surface phenomena and cannot be naturally generated in deep structure. Phrase structure grammars for L_1 and L_2 can be combined to form a code-switching grammar which generates grammatical monolingual sentences as well as those containing only valid code-switches.

Turning to the data analysis itself, we find that rule probabilities for the code-switching grammar represent a compromise between G_1 and G_2 probabilities, but the details of this compromise remain to be investigated. Finally, the switching propensities for various syntactic boundaries yield a clear and simple picture of syntactic

effects on code-switching. For most boundary types, switches occur with a rate proportional to the occurrence of the boundary type. Freely moveable constituents have more switches at their boundaries, while boundaries between constituents which are highly constrained to occur together, approaching the status of bound morphemes, are more resistant to switches.

We do not claim perfect accuracy for all the figures in Table 2, given the size of our sample, possibilities of incompatibility of the two corpora used, and the rough nature of the syntactic analysis. Nevertheless, their interpretation is quite clear. The equivalence and free morpheme constraints extend quantitatively to performance data: not only are all boundaries which satisfy the equivalence constraint eligible for code-switching, but most are equally likely to be the site of a switch. Those exceptional boundaries which show a relatively low rate of switching involve two closely bound syntactic elements whose relationship approaches, but does not quite enter, the domain of the free morpheme constraint. This quantitative approach permits an analysis which accounts for more of the data and is more scientific than the constraint-and-exception paradigm which has characterized the code-switching literature.

To the extent that the code-switching constraints, both in their qualitative and quantitative aspects, are validated by this and future studies, they may prove to be useful tools in the study

of monolingual syntactic structure. We have already seen, for example, that the free morpheme constraint prohibits switches categorically only between truly bound forms, but that it operates in a weaker way between forms which are closely linked but not clearly bound. We may now reverse the argumentation and make use of this fact to evaluate the status of binding relationships between morphemes in monolingual speech. If two supposedly bound morphemes in a language are investigated in a code-switching situation and found never to be separated by a code-switch, their bound status is confirmed. If their boundary is susceptible to switches, but only at a low rate, we may say they are weakly bound, and so on.

Similarly, for the equivalence constraint, where there is some question over the rules generating a certain class of structures in monolingual speech, an investigation of the proposed syntactic boundaries in the code-switching situation may help clarify the situation. For example, Spanish preposed objects may be generated in two ways: directly in the verb phrase, as in (27), or by topicalizing extraposition, as in (28).

(27) VP → NP V

Ellos al gato mataron. 'They killed the cat.'

(28) S → NP S

Al gato, ellos mataron. 'The cat, they killed.'

Since subject pronoun deletion is common in Spanish, both (27) and (28) reduce to (29):

(29) *Al gato mataron.* 'They killed the cat.'

For a speech variety where sentences like (29) are common (not the case for Puerto Rican Spanish), an investigation of the possibility of switches into English between *gato* and *mataron* would be diagnostic of the syntactic structure. If (29) has the same structure as (28), such switches would be common. If the structure is like (27), they would be prohibited, since English cannot prepose object NPs in the VP.

A third area where code-switching may be an indication of syntactic structure is in evaluating the relative importance of constituent hierarchy and lexicon in the structure of sentences. For example, prepositions (or subordinate conjunctions) introducing a verb complement may be heavily constrained lexically, i.e. by the verb in question. In the constituent hierarchy, however, those items will be more closely grouped with the other elements of the complement than with the verb. The possibility that switches occur more readily after prepositions and subordinate conjunctions than before, may reflect the greater weight of lexicon-controlled constraints than constituent hierarchy relationships. This may be especially true if a verb required different complement structures in the two languages.

The evidence we have presented for the syntactic integrity of Spanish and English grammars, even when they are being used sequentially and simultaneously, bolsters other arguments for non-

convergence of Spanish and English in the Puerto Rican speech community. A quantitative semantic analysis of tense and aspect (Pousada and Poplack 1979) and morphophonological analysis of word-final inflections (Poplack 1980) in the same community have also shown that the grammar of Spanish (aside from the lexicon), which serves a wide range of communicative functions, has been extraordinarily resistant to influence from the grammar of English; this despite the economic and political dominance of the English-speaking community.

This integrity of the monolingual modes of discourse in the community clearly puts into relief the special nature of the code-switching mode as a distinct communicative resource for skilled bilingual speakers. This mode, which is not to be confused with borrowing or other language contact phenomena, is governed by a well-defined set of syntactic rules. We have shown its structure to be accessible through the scientific study of speech performance in much the same way as monolingual varieties.

Notes

[1] This research was supported by an operating grant from the Natural Sciences and Engineering Council of Canada and by grants from the National Institute of Education (U.S.A.) and the Ford Foundation. Part of this paper was presented at the 54th Meeting of the Linguistic Society of America (1979). Many of the points raised here reflect discussions with Alberto Rivas, Claire Lefebvre, Gillian Sankoff, and members of the Language Policy Task Force at the Center for Puerto Rican Studies. Thanks to Migdalia Rodriguez for her expert preparation of the text.

[2] The situation with trilingual code-switching gives rise to analogous questions, somewhat more complicated.

[3] This includes the transformational grammar approach of Barkin and Rivas (1979), despite their concern for keeping the monolingual grammars separate while generating the set of code-switched sentences.

[4] The difficulty in constructing an example with less shaky postulates is a consequence of the shrinking stock of transformations now recognized by generative grammarians.

[5] This may seem like an uneconomical procedure. Why put $^{sp:v}$ superscripts on rule output categories which never dominate verbs? Would it not be preferable to limit the number of categories in the code-switching grammar distinguished only by irrelevant superscripts? The answer is yes--for any particular code-switching grammar. It is a simple matter to determine which categories can dominate a V, which an N, and so on, in English and Spanish. But for an arbitrary code-switching grammar, this means devising an algorithm to determine exactly which non-terminal categories may dominate which terminal categories in a potential equivalence constraint violation. This should not be difficult, and may well be preferable, but to keep the present already complicated exposition as short as possible, we omit the discussion of such an algorithm, at the expense of a proliferation of superscripts.

[6] For example, even in the present case of Spanish-English code-switching, it seems probable that it will be necessary to include certain 'hybrid' rules. Here the first half of the rule output will reflect a strictly Spanish pattern, say, while the second half will be purely English, but there is no constraint against

switching between the two halves. This is a complication in detail only, and we will not discuss it further here.

[7] Certain boundary types appear collapsed in the Table, e.g. the four combinations between N or NP and V or VP, because of differences in the boundary frequency calculations and the coding of the original data; although only NP VP boundaries are generated by our code-switching grammar, some switches had previously been coded N VP, N V or NP V.

References

Barkin, F., and A. Rivas. 1979. On the underlying structure of bilingual sentences. Paper presented at the Linguistic Society of America, 54th Annual Meeting, Los Angeles.

Denison, N. 1971. Some Observations on Language Variety and Plurilingualism. In J.B. Pride and J. Holmes (eds.) Sociolinguistics. Middlesex: Penguin. 65-77.

Gal, S. 1978. Variation and Change in Patterns of Speaking: Language Shift in Australia. In D. Sankoff (ed.) Linguistic Variation: Models and Methods. New York: Academic Press. 227-238.

Gingràs, R. 1974. Problems in the Description of Spanish-English Intra-sentential Code-switching. In G.A. Bills (ed.) Southwest Areal Linguistics. San Diego: Institute for Cultural Pluralism.

Granda, G. de. 1968. Transculturación e interferencia lingüística en el Puerto Rico contemporáneo (1898-1968). Bogotà: Instituto Caro y Cuervo.

Grenander, U. 1967. Syntax-controlled probabilities. Technical Report, Brown University, Division of Applied Mathematics.

Gumperz, J. 1964. Linguistic and Social Interaction in Two Communities. In J. Gumperz (ed.) Language in Social Groups. Stanford: Stanford University Press. 151-176.

_____. 1971. Bilingualism, Bidialectalism and Classroom Interaction. In Language in Social Groups. Stanford: Stanford University Press. 311-339.

_____. 1976. The sociolinguistic significance of conversational code-switching. Working papers of the Language Behavior Research Laboratory #46. Berkeley: University of California.

_____., and E. Hernández Chavez. 1970. Cognitive Aspects of Bilingual Communication. In E. Hernández-Chavez et al. (eds.) El lenguaje de los Chicanos. Arlington: Center for Applied Linguistics. 154-163

_____., and R. Wilson. 1971. Convergence and Creolization: A Case from the Indo-Aryan/Dravidian Border. In D. Hymes (ed.)

Pidginization and Creolization. London: Cambridge University Press. 151-167.

Hasselmo, N. 1972. Code-switching as Ordered Selection. In E. Finchow et al. (eds.) Studies for Einar Haugen. The Hague: Mouton.

_____. 1979. Code-switching and Modes of Speaking. In G. Gilbert (ed.) Texas Studies in Bilingualism. Berlin: Walter de Gruyter and Co.

Heidelberger Forschungsprojekt "Pidgin-Deutsch". 1978. The Acquisition of German Syntax by Foreign Migrant Workers. In D. Sankoff (ed.) Linguistic Variation: Models and Methods. New York: Academic Press. 1-22.

Hindle, D. 1979. A probabilistic grammar of noun phrases in English. Paper presented at NWAVE 8th Annual Meeting, Montreal.

Horning, J.J. 1969. A study of grammatical inference. Technical Report No. CS 139, Computer Science Department, Stanford University.

Klein, S. 1965. Control of style and generative grammar. Language 41. 619-631

Klein, W., and N. Dittmar. 1979. Developing grammars. Berlin: Springer-Verlag.

LaFontaine, H. 1975. Bilingual education for Puerto Ricans: ¿si o no? Paper presented at the National Conference on the Educational Needs of the Puerto Rican in the United States. Cleveland, Ohio.

Lance, D. 1975. Spanish-English Code-switching. In E. Hernández-Chavez et al. (eds.) El lenguaje de los Chicanos. Arlington: Center for Applied Linguistics. 138-153.

Lipski, J. 1978. Code-switching and the Problem of Bilingual Competence. In M. Paradis (ed.) Aspects of Bilingualism. Columbia, S.C.: Hornbeam Press. 250-264.

McClure, E. 1977. Aspects of code-switching in the discourse of bilingual Mexican American children. Washington, D.C.: Georgetown University Rountable on Languages and Linguistics.

_____., and J. Wentz. 1975. Functions of code-switching among Mexican American children. In Papers from the Parasession on Functionalism. Chicago: Chicago Linguistic Society. 421-432.

Pedraza, P. ms. Ethnographic observations of language use in El Barrio.

Pfaff, C. 1975. Syntactic constraints on code-switching: a quantitative study of Spanish/English. Paper presented at the Linguistic Society of America, 50th Annual Meeting.

_____. 1976. Functional and structural constraints on syntactic variation in code-switching. Papers from the Parasession on Diachronic Syntax. Chicago: Chicago Linguistic Society.

_____. 1979. Constraints on language mixing. Language 55. 291-318.

Poplack, S. 1978. Syntactic Structure and Social Function of Code-switching. In R. Durán (ed.) Latino Discourse and Communicative Behavior. New Jersey: Ablex Publishing Corp. (in press).

_____. 1979a. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL. Working Paper #4. New York: Center for Puerto Rican Studies.

_____. 1979b. Consideraciones tipológicas en la alternancia de lenguas. Proceedings of the 4th Annual Symposium on Caribbean Dialectology. Interamerican University Press. (to appear).

_____. 1980. Variable concord in sentential plural marking. Working Paper #6. New York: Center for Puerto Rican Studies.

Pousada, A., and S. Poplack. 1979. No Case for Convergence: the Puerto Rican Spanish Verb System in a Language Contact Situation. In J. Fishman and G. Keller (eds.) Bilingual Education for Hispanic Students in the United States. Columbia University: Teachers College Press. (to appear).

Rayfield, J.R. 1970. The languages of a bilingual community. The Hague: Mouton.

Sankoff, D. 1971. Branching processes with terminal types: application to context-free grammars. Journal of Applied Probability 8. 233-240.

_____. 1972. Context-free grammars and nonnegative matrices. Linear Algebra and its Application 5. 277-281

Sankoff, G. 1968. Social aspects of multilingualism in New Guinea. Ph.D. dissertation. McGill University.

_____. 1972. Language Use in Multilingual Societies: Some Alternative Approaches. In J.B. Pride and J. Holmes (eds.) Sociolinguistics. Middlesex: Penguin. 33-51.

Sciullo, A.-M. di, A. van Ameringen, H.J. Cedergren, and P. Pupier. 1976. Etude de l'interaction verbale chez des Montréalais d'origine italienne. Cahiers de Linguistique de l'Universite du Québec 6. 127-153.

Suppes, P. 1970. Probabilistic grammars for natural languages. Synthèse 22. 95-116.

Timm, L.A. 1975. Spanish-English code-switching: el porqué y how-not-to. Romance Philology 28, #4. 473-482.

_____. 1978. Code-switching in War and Peace. In M. Paradis (ed.) Aspects of Bilingualism. Columbia, S.C.: Hornbeam Press. 302-315

Valdés-Fallis, G. 1976. Social Interaction and Code-switching patterns: A Case Study of Spanish/English Alternation. In G. Keller et al. (eds.) Bilingualism in the Bicentennial and Beyond. New York: Bilingual Press. 53-85

_____. 1978. Code-switching as a Deliberate Verbal Strategy: A Microanalysis of Direct and Indirect Requests Among Bilingual Chicano Speakers. In R. Durán (ed.) Latino Discourse and Communicative Behavior. New Jersey: Ablex Publishing Corp. (in press).

Varo, C. Dr. 1971. Consideraciones antropológicas y políticas en torno a la enseñanza del 'Spanglish' en Nueva York. Río Piedras: Ediciones Librería Internacional.

Weinreich, U. 1953. Languages in Contact. The Hague: Mouton.

Wentz, J. 1977. Some considerations in the development of a syntactic description of code-switching. Ph.D. dissertation. University of Illinois at Urbana-Champaign.